Public Grievances of Indian Railways: An analysis of Twitter data

From: Sai Srujana Illa
To: Professor Brodnax
Date: December 14, 2022

Executive Summary

Under the Modi government, the Indian Railways implemented many customer-central reforms to improve service delivery. The Modi government has time and again reiterated its focus on minimum government and maximum governance, one of the prominent promises made by BJP in their 2014 election manifesto (BJP, 2014). The concept of minimum government and maximum governance aims at improving efficiency in public sector operations through administrative reforms which bring about accountability and transparency (Bhattacharya, 2021).

One can view transparency as a tool to improve efficiency and to motivate reforms. Literature shows mixed results for the impact of transparency and accountability on governance(Bannister & Connolly, 2011; Kosack & Fung, 2014). An Indonesian study on the effectiveness of internal and external accountabilities on performance of local government officials shows that while internal accountability had a positive influence on performance of employees, external accountability did not (Risakotta & Akbar, 2019). Thus, the literature on the effect of public accountability on bureaucratic performance is mixed. Nonetheless, public availability of data plays a crucial role in informing public opinion and thus reforms.

While the Modi government has brought in several reforms such as e-Office, e-Samiksha, Citizen Charters etc. to improve service delivery timeframe, transparency, and internal accountability, public accountability has been neglected. Public availability of metrics on the operations of government run sectors is at most minimal. Data is made public only when it is requested through a Right to Information Act (RTI) filing or as a response to questions raised in the parliament(Livemint, 2021; News18, 2021). In this light, where the official Indian Railways grievance data is not publicly available, we analyze the public grievances made to the Indian Railways through Twitter.

On classifying and analyzing public grievance tweets, the categories that stood out in terms of the sheer volume of grievances are the ones related to railway exams/jobs, cleanliness, refunds, and reservations. These insights from the analysis highlight well-known issues of railway recruitment and operations. The results show that the Indian Railways needs to focus on their emergency ticket booking policy and should expand ‘Swachh Rail’ initiative to identified trains with higher number of cleanliness grievances. An area which has not been of much focus is the impact of sporadic railway recruitment on the productivity loss for the country, with regards to the youth waiting either for railway entrance exams to be conducted or the results to be declared.

This study involved the classification of public grievance tweets based on sentiment analysis and further categorization using topic modeling. Given that public grievances to Indian Railways are made through various platforms such as ‘Centralised PublicGrievance Redress and Monitoring System’ (CPGRAMS), RailMadad, Twitter, Facebook, Instagram and Youtube, the distribution of public grievances might be subject to change if data from all these sources is considered.

Introduction and Background

The functioning of Indian government departments is highly scrutinized for high lead times in addressing grievances. In 2021, Indian Railways has launched an application called Rail Madad, a one-stop solution for Inquiry, Assistance & Grievance Redressal (RailMadad, 2021). However, data on the complaints is not made public thus neglecting public accountability. In the meantime, the microblogging platform twitter has evolved into a space where citizens share their grievances. In 2017, Indian railways has set up its official public grievances account RailwaySeva (@RailwaySeva, 2017). However, public grievance tweets primarily

mention only the account of the Indian Railway Ministry (@RailMinIndia, 2014). But the mentions of the Railway Ministry have many kinds of tweets other than public grievances on their operations.
In 2022, the Indian Railways recorded the 4th highest (54k) number of public grievances on CPGRAMS (CPGRAMS, 2022), the central public grievance system for all government run statutory bodies. However, the complaints registered on CPGRAMS, based on our calculations are only an estimated 2.2% of the grievances received on Twitter (6500 per day)(Indian Railways, 2016). Twitter has indeed evolved as an important platform for public grievances.
Therefore, in this project we use text analysis methods such as sentiment analysis and topic modelling on twitter data with the following goals: i) To assess whether public grievance tweets can be classified from non-public grievance tweets using sentiment analysis ii) Classification of the public grievances tweets under different categories and analysing the distribution of public grievances by categories and identification of the pain points

Literature Review

Many studies infer the gender of twitter users based on various content available on their Twitter profiles; user names, profile pictures, description and contents of their tweets (Fink et al., 2012; Yang et al., 2021). In this study, we infer gender simply based on name using the python package GuessIndianName (v-adhithyan, 2019). While using just the username is not the best strategy given that twitter users can use any name as their username, it is a good place to start with.

A similar work by Akhtar et al. uses SVM classifiers for the classification of public grievance tweets (Akhtar & Beg, 2021). Another study on complaint detection in social media compared between various methods, showing that semantic features and sentiment features show high predictive accuracy in this context (Gautam, 2020). Sentiment Analysis is widely used on social media, survey responses and reviews data to identify subjective tone of the domain specific text. In the current study we opt for unsupervised learning such as sentiment analysis and topic modelling. Topic Modelling is a technique used to identify underlying topics in text and is widely used for the classification of documents. We choose these methods to circumvent hand coding of tweets. However, the validation of results will be based on hand coding.

Data Collection:

We collect tweets by utilizing tweepy (Roesslein, 2009) a python library used for connecting to the twitter API (Twitter, a). With a developer account we used the query '@RailMinIndia -filter:retweets' using search_tweets method to extract tweets which tagged the main twitter handle of the Ministry of Railways (@RailMinIndia, 2014) and the ones which are not retweets. The language parameter to English to pull only those tweets which are in English. To get the full text of the tweet instead of the truncated version, the 'extended' tweet mode is used while extracting the data. It returns a json object, whose full text property has the untruncated version of the tweet. Our final dataset has 18K tweets which were tweeted over a span of one week. Table 1 shows the desired features which are extracted using the API.

Features	Description	Missingness (%)
Tweet ID	ID of the tweet which mentions the twitter	0
Tweet text	Text content of the tweet	0
Created at	Time when the tweet was made	0
Username	Name of the user who made the tweet	0
User location	Location of the user when the tweet was made.	44

Table 1: List of variables and their percentage missingness in the dataset

The tweet ID field is used in identifying and removing duplicate tweets. The 'created at' field is used to analyse the traffic of grievances with respect to various categories at different times of the day. User location is used to analyse which regions have more operational issues.

- Tweets might have hashtags (starting with #), mentions (starting with @), and links (starting with http or www). Regex is used to identify and remove such sequence of characters.
- We substitute all non-alphanumerical characters in the tweets with space and then reduce the spacing between spaces to single spaces. With this pre-processing if a tweet was of another language, we will be able to identify it as it will have very few tokens.
- The next pre-processing step is to remove stop words and perform vectorization. We use count vectorizer to build the document term matrix, where the value for each token is equal to the number of times it has appeared in a particular tweet. We then analyse the number of words in each tweet by creating a field which is the sum of all the word counts. Figure 1 shows the distribution of number of words in the tweets. The outlier tweets are then deleted as they are not informative.

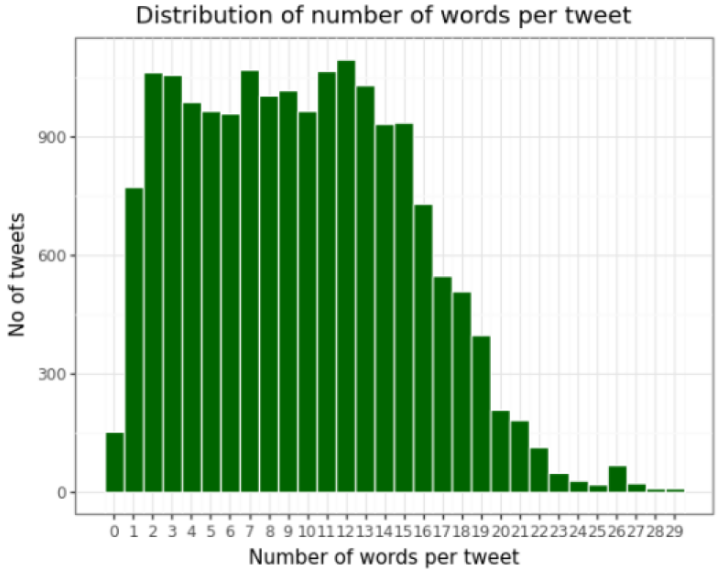

Figure 1: Distribution of the count of words in the tweets

- As we see in Figure 2 we see that there are some context specific stop words such as 'train', 'railway', 'pnr' which are further removed. There are also plural and grammatical variants of the same words being mentioned. To deal with this redundancy, we perform stemming or lemmatization where words are converted to their root form. Figure 3 shows the top words after pre-processing.

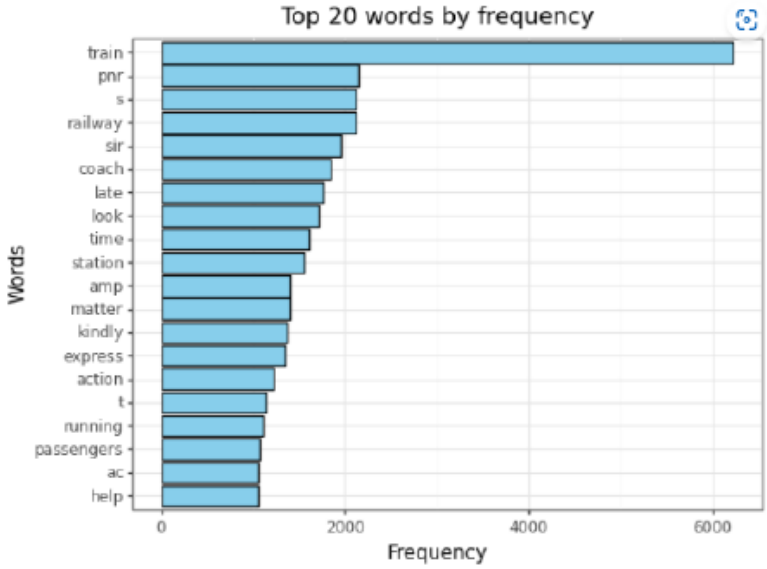

Figure 2: Top 20 words by frequency before pre-processing

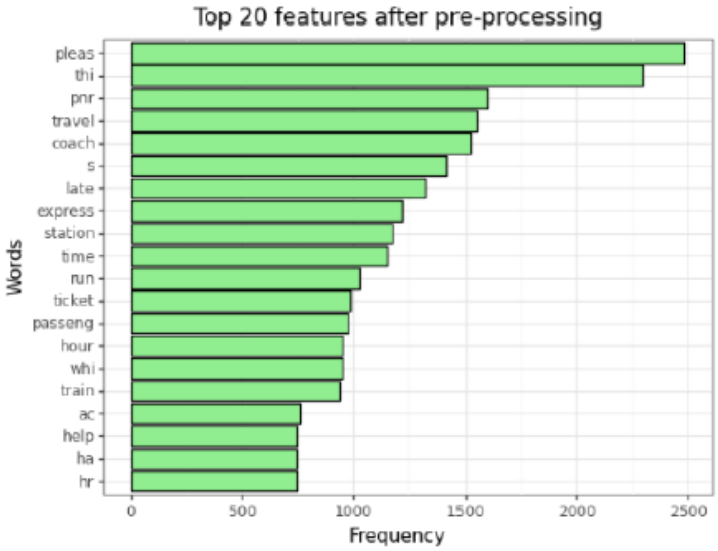

Figure 3: Top 20 words by frequency after pre-processing

- From Figure 2 and Figure 3, we see names of locations in our feature set. This could be useful information which can be used as a proxy for the complainant location at the time of their complaint. Given that the location field has 44% missing values, we could use locations identified through this method as a proxy. We use Spacy's Named Entity Recognition (NER) to extract these locations(Li, 2018).
- We then perform TF-IDF vectorization without scaling but with normalization, which assigns the relative importance of token based on their occurrence in a particular tweet and across the tweets. We do not perform scaling as tweets are concise in nature, but we perform normalization due to varied length of tweets.
- From Figure 2, we select a minimum document frequency of 45, which results in a feature set of 622 features.

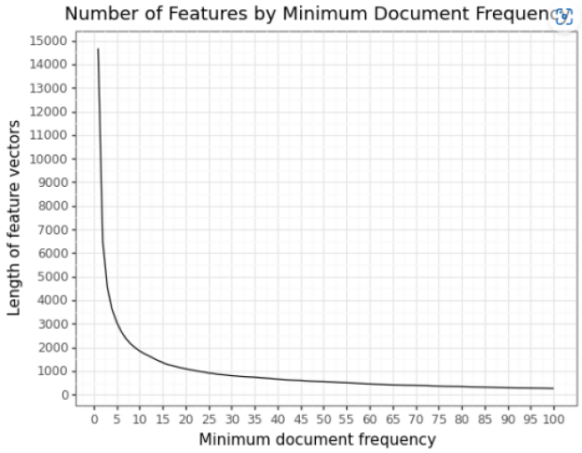

Figure 4: Number of features by minimum document frequency

Methodology

We use various text analytics techniques both in preparation of the dataset and analyzing it. In dataset preparation, it will be interesting to have the location from where tweets are made, as it'll help to gauge the intensity of grievances across various railway zones. However, on extracting location information from the twitter API, we observe that there is high level of missingness (around 44%). A workaround to this problem is to extract location from the text using "Named Entity Recognition" (NER) technique. NER technique involves tagging parts of speech in a sentence and thereby identifying entities and further classifying them

under categories such as ‘Name’, ‘Place’, ‘Organization’ etc. We use NER method from the NLTK package to extract location information.

In India, Twitter usage is disproportionate by gender. While 70% of India Twitter users are males, only 30% are females (Vardhman, 2021). It could be the case that people of the opposite sexes are differentially disappointed within the various operational categories of the Indian Railways and their usage of twitter for public grievance is varied. Therefore, it could be that the case that our results underestimate certain public grievance categories give the demographic distribution of twitter users. Therefore, gender could be an important variable for our analysis. We infer gender based on username using the GuessIndianName package. In addition to the accuracy rate of this package and another limitation is that twitter usernames are not necessarily the names of the users.

The first goal of this project is to assess whether public grievances can be classified using sentiment analysis tweets as to whether they are public grievances. As we do not have labelled data, we go for sentiment analysis which is an unsupervised learning technique. Sentiment analysis will be performed before vectorization of the tweets. Sentiment analysis traditionally depends on a dictionary of important words with scores assigned to them. These scores represent the positive or negative nature of the words. Classification is then performed based on the words frequency weighted by the scores. However, the dictionary method regards words to be independent and therefore information in the sentence strucre is lost. We use Bidirectional Encoder Representations from Transformers (BERT) model a attention mechanism that learns contextual relations between words (Horev, 2018). However, a trade-off is that dictionary model is more explainable compared to BERT. We will further validate the results by choosing a random sample of tweets and comparing the results against hand coded labels.

The second goal of this study is to categorise the public grievances into different categories. To identify the latent categories in the tweets corpus I will use topic modelling based on Latent Dirichlet Allocation (LDA) approach, where the probabilities of a tweet belonging to a predetermined number of topics can be calculated. We choose this soft clustering technique, as we see that passengers tweet on many issues at the same time. The LDA approach initially makes assumptions on the per-document topic distribution and per-topic word distribution (Doll, 2019). Based on these assumptions, fake documents are generated which are then compared to the real document. By maximizing the likelihood that the fake documents could have generated the real documents, we arrive at a distribution of topics per document and distribution of words per topic.
After getting the categories of the grievances through topic modelling, the dataset will then be ready for further analysis.

Findings

Figure 5 shows the results of gender inference based on usernames. This distribution is even more unequal compared to the 70-30 gender distribution of Twitter users (Vardhman, 2021).

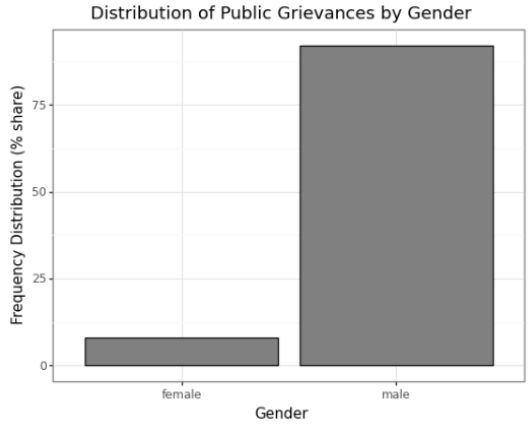

Table 2 shows the results of sentiment analysis for the classification of public grievances. Validation is performed on a random sample of 100 tweets. We see that the accuracy rate is only 80%. These results could be further improved by complementing the sentiment analysis results by other methods like n-gram search or supervised classification methods.

	True Negative	True Positive
Predicted Negative	68	8
Predicted Positive	12	12

Table 2: Results of public grievance classification by sentiment analysis

From Figure 6 we see that more than 15% of public grievances are related to train delays. The topic 'Sleeper & General coaches' covers the issue of individuals with general tickets occupying the sleeper coaches. A general public opinion is that the number of sleeper coaches in trains needs to be increased given that there is high demand for sleeper coaches. Public grievances related to reservations/ refunds and cleanliness are other commonly observed categories. An interesting topic that we see in our results is that of jobs and exam results. This topic is related to the public grievances of those youngsters who intend to write the Railways entrance examination or who are waiting for the results.

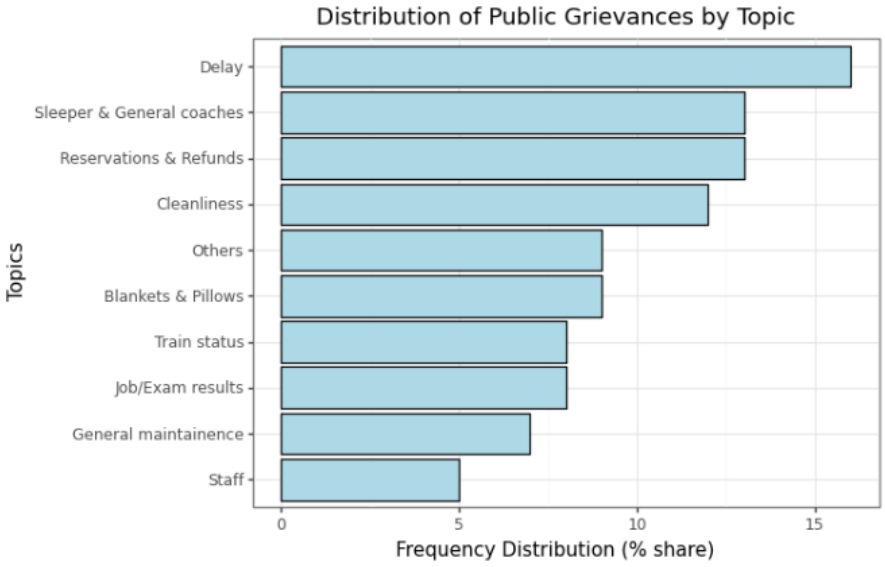

Figure 6: Distribution of Public Grievances across categories

Implications & Policy Recommendations

Indians' affinity for government jobs is driven by the job security that government jobs offer. It is a well-known issue that the examinations for these jobs are conducted sporadically. This leads to many young Indians spending their precious time preparing for exams or waiting for results. Literature shows that this waiting leads to productivity loss and leads to suboptimal economic outcomes for the youth involved in the long run (Mangal, 2021). Mangal notes that "as long as candidates face strong incentives to continue studying, a regular and timely testing policy will help reduce the unemployment rate"(Mangal, 2021). But in the long run it would be ideal to shift to an alternative recruitment process which places lower emphasis on the candidates' exam score, in order to disincentivize longer periods of unemployment.

Another salient feature of the Indian Railway operations which had high number of complaints on is the Tathkal reservation system, wherein some tickets are made available two days prior to the travelling date and people either get reserved seats or can choose to be waitlisted. Many passengers show up at the railway stations with their waitlisted tatkal reservations to try their luck and are disappointed with their seat allocation. Given that all the tweets do not have train number, if we use user location as a proxy, trains passing through Mumbai, Delhi and Hyderabad have the highest number of tatkal related grievances. The Indian Railways should monitor the demand on these train routes and consider expanding capacity or provide an alternate booking policy or with a better refund system. If analysis of tweets is to be adopted for this monitoring purpose, it would be ideal to have the train number on tweets. Consumers could be nudged to enter train information by having autocomplete statements or prompts when writing tweets which tag the Indian Railway accounts.

With regards to clean maintenance of trains and stations, Indian Railways has launched Swachh Rail program wherein certain trains has been equipped with bio-toilets and 'Clean my Coach' service. By tracking the public grievance categories and details as in this study, Indian Railways will further be able to identify trains or stations which need attention.

Conclusion & Limitations

Sentiment analysis is not very effective in classifying public grievances as they are not always in a negative tone. we need additional methods such as n-gram search to detect grievances which are positive in tone. We also observe that the LDA topic modelling results are not well separated. This could be also because of nature of these short tweets, where multiple topics are discussed. For instance, a person might have a complaint about how unclean the bedsheets and pillows are and this tweet who cover two topics (Ho, 2018). Therefore, other clustering methods need to be explored for better results.

Given that public grievances to Indian Railways are made through various platforms, the distribution of public grievances might be subject to change if data from all these sources is considered. Given the disproportionate usage of Twitter across various demographic variables in India, it could be that the case that our results underestimate certain public grievance categories. Tweets in Regional Languages are discarded, assuming that they are distributed at random over the grievance categories. If not, the distribution of public grievances across categories will be subject to change.

Bibliography

Akhtar, N., & Beg, M. M. S. (2021). Railway Complaint Tweets Identification. In N. Sharma, A. Chakrabarti, V. E. Balas, & J. Martinovic (Eds.), *Data Management, Analytics and Innovation* (pp. 195–207). Springer. https://doi.org/10.1007/978-981-15-5616-6_14

Bannister, F., & Connolly, R. (2011). The Trouble with Transparency: A Critical Review of Openness in e-Government. *Policy & Internet*, *3*(1), 1–30. https://doi.org/10.2202/1944-2866.1076

Bhattacharya, D. S. (2021). *ANALYZING THE CONCEPT OF MINIMUM GOVERNMENT AND MAXIMUM GOVERNANCE*. *10*(3).

BJP. (2014). *Bharatiya Janata Party*. Bharatiya Janata Party. https://www.bjp.org/

CPGRAMS. (2022). *CPGRAMS Dashboard*. https://pgportal.gov.in/darpgdashboard

Fink, C., Kopecky, J., & Morawski, M. (2012). Inferring Gender from the Content of Tweets: A Region Specific Example. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), Article 1. https://doi.org/10.1609/icwsm.v6i1.14320

Gautam, A. K. (2020, July 7). *Identification of Complaint Relevant Posts on Social Media*. Medium. https://towardsdatascience.com/identification-of-complaint-relevant-posts-on-social-media-4bc2c8b625ca

Ho, G. (2018, March 6). *Why Latent Dirichlet Allocation Sucks*. ⁂ George Ho. https://www.georgeho.org/lda-sucks/

Horev, R. (2018, November 17). *BERT Explained: State of the art language model for NLP*. Medium. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Indian Railways. (2016). *Presentation on Achievements & Plans of Indian Railways*. https://indianrailways.gov.in/IndianRailways/Presentation%20on%20Achievements%20&%20%20Plans%20of%20Indian%20Railways.pdf

Kosack, S., & Fung, A. (2014). Does Transparency Improve Governance? *Annual Review of Political Science*, *17*(1), 65–87. https://doi.org/10.1146/annurev-polisci-032210-144356

Li, S. (2018, December 6). *Named Entity Recognition with NLTK and SpaCy*. Medium. https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da

Livemint. (2021). *Indian Railways one-stop passengers' helpline "Rail Madad" launched*. Mint. https://www.livemint.com/news/india/indian-railways-one-stop-passengers-helpline-rail-madad-launched-11628299751201.html

Mangal, K. (2021). *Competitive Exams for Government Jobs and the Labor Supply of College Graduates in India*. https://kmangal.github.io/files/papers/competitive-exams.pdf

News18. (2021). *Indian Railways' Grievance Portal Got Over 5,500 Complaints from Telangana This Year, Reveals RTI*. News18. https://www.news18.com/news/india/indian-railways-grievance-portal-got-over-5500-complaints-from-telangana-this-year-reveals-rti-4322762.html

RailMadad, A Grievance Redressal Mechanism. (n.d.). Retrieved October 4, 2022, from https://railmadad.indianrailways.gov.in/madad/final/home.jsp

@RailMinIndia. (2014). *(2) Tweets with replies by Ministry of Railways (@RailMinIndia) / Twitter*. Twitter. https://twitter.com/RailMinIndia

@RailwaySeva. (2017). *(@RailwaySeva) / Twitter*. Twitter. https://twitter.com/RailwaySeva

Risakotta, K., & Akbar, R. (2019). THE EFFECT OF INTERNAL AND EXTERNAL ACCOUNTABILITY, JOB MOTIVATION AND EDUCATION ON LOCAL GOVERNMENT OFFICIAL'S PERFORMANCE. *Journal of Indonesian Economy and Business*, *33*, 257. https://doi.org/10.22146/jieb.13921

Roesslein, J. (2009). *Tweepy*. https://www.tweepy.org/

Twitter. (n.d.). *Twitter API Documentation*. Retrieved October 4, 2022, from https://developer.twitter.com/en/docs/twitter-api

v-adhithyan. (2019). *guess-indian-gender: Guess gender from indian names* (1.0.1) [Python; OS Independent]. https://github.com/v-adhithyan/gender-guess-indiannames

Vardhman, R. (2021). *Twitter Users in India—(Statistics & Facts) | 2022*. https://findly.in/twitter-users-in-india-statistics/

Yang, Y.-C., Al-Garadi, M. A., Love, J. S., Perrone, J., & Sarker, A. (2021). Automatic gender detection in Twitter profiles for health-related cohort studies. *JAMIA Open*, *4*(2), ooab042. https://doi.org/10.1093/jamiaopen/ooab042